Journal of Biomedical Informatics xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Data mining methods for classification of Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data

Tim Van den Bulcke [a,*], Paul Vanden Broucke [a], Viviane Van Hoof [b,c,d], Kristien Wouters [a], Seppe Vanden Broucke [a], Geert Smits [a], Elke Smits [a], Sam Proesmans [d], Toon Van Genechten [d], François Eyskens [b,d,e]

[a] i-ICT, University Hospital Antwerp, Wilrijkstraat 10, 2650 Edegem, Belgium
[b] Provinciaal Centrum voor de Opsporing van Metabole Aandoeningen (PCMA), Antwerpen, Belgium
[c] Dept. of Clinical Chemistry, University Hospital Antwerp, Edegem, Belgium
[d] Faculty of Medicine, University Antwerp, Antwerpen, Belgium
[e] Dept. of Paediatrics/Metabolic Diseases, University Hospital Antwerp, Edegem, Belgium

ARTICLE INFO

Article history:
Received 15 July 2010
Available online xxxx

Keywords:
Data mining
Rare diseases
MCADD
Medium-Chain Acyl-CoA dehydrogenase
Logistic regression

ABSTRACT

Newborn screening programs for severe metabolic disorders using tandem mass spectrometry are widely used. Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) is the most prevalent mitochondrial fatty acid oxidation defect (1:15,000 newborns) and it has been proven that early detection of this metabolic disease decreases mortality and improves the outcome. In previous studies, data mining methods on derivatized tandem MS datasets have shown high classification accuracies. However, no machine learning methods currently have been applied to datasets based on non-derivatized screening methods.

A dataset with 44,159 blood samples was collected using a non-derivatized screening method as part of a systematic newborn screening by the PCMA screening center (Belgium). Twelve MCADD cases were present in this partially MCADD-enriched dataset. We extended three data mining methods, namely C4.5 decision trees, logistic regression and ridge logistic regression, with a parameter and threshold optimization method and evaluated their applicability as a diagnostic support tool. Within a stratified cross-validation setting, a grid search was performed for each model for a wide range of model parameters, included variables and classification thresholds.

The best performing model used ridge logistic regression and achieved a sensitivity of 100%, a specificity of 99.987% and a positive predictive value of 32% (recalibrated for a real population), obtained in a stratified cross-validation setting. These results were further validated on an independent test set. Using a method that combines ridge logistic regression with variable selection and threshold optimization, a significantly improved performance was achieved compared to the current state-of-the-art for derivatized data, while retaining more interpretability and requiring less variables. The results indicate the potential value of data mining methods as a diagnostic support tool.

1. Introduction

The early diagnosis of rare diseases constitutes a great challenge in current medicine. Currently rare diseases are often diagnosed too late, resulting in a decline in life expectancy and quality of life, and an increase in healthcare costs. The large variety of rare diseases makes that – although each disease affects only a small number of people – it still affects a large population. As such, the total number of patients suffering from a rare disease in Europe is around 30 million [1]. Early detection and diagnosis of metabolic disorders and of rare diseases in general, are of crucial importance for the further outcome of the patient. As such, statistical and machine learning methods could be of great value as a diagnostic support tool for doctors and medical personnel.

1.1. MCADD

This study focuses on MCADD (Medium-Chain Acyl-CoA dehydrogenase deficiency), the most frequent metabolic disorder of mitochondrial fatty acid oxidation [2]. There are four categories of fatty acids, differentiated by their carbon chain length: short

* Corresponding author.
E-mail address: tim.van.den.bulcke@uza.be (T. Van den Bulcke).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.12.001

(C2–4), medium (C4–12), long (C12–20) and very long (>C20). The enzymes responsible for their β-oxidation are respectively SCAD, MCAD, LCAD and VLCAD (small, medium, long and very long chain acyl-Co-A dehydrogenase).

Our body preferentially metabolizes carbohydrates, but only a limited stock of carbohydrates is available and after a fasting period of 9–10 h (e.g. a normal night of sleep), the body switches to energy production from fatty acids. Defects of mitochondrial fatty acid β-oxidation therefore lead to a disturbed or inhibited energy production of fatty acids. Inherited defects fall into three groups: (a) those associated with the carnitine mediated transport into the mitochondria; (b) those of the matrix enzymes (such as MCAD); and (c) those affecting the activity of membrane-bound enzymes of long-chain fatty acid oxidation.

In case of MCADD, the production of the MCAD enzyme is absent or reduced. As such, the β-oxidation of the fatty acids C4 and higher fails and they can subsequently not be used as an energy source. The symptomatic MCADD-patient shows a clinical picture strongly resembling Reye's syndrome with hepatomegaly and stupor associated with hypoketonemia, hypoglycemia, hypocarnitinemia, increased transaminsase and mild hyperammonemia [3]. Lipids that cannot be used precipitate in liver, heart, kidneys. These patients present themselves over the years with hepatomegaly or hepatic steatosis, cardiomyopathy, encephalopathy and decreased muscle tone. 10–20% of the patients develop rhabdomyolysis in the first three years of life, even when adequately treated [4].

The early diagnosis of MCADD – and metabolic diseases in general – is crucial for the further outcome and prognosis of the patient. If the diagnosis is made early, the quality of life can be substantially improved. With supplementation of acylcarnitine and a diet high in carbohydrates and low in fats and fasting periods not longer than 6 h, the prognosis for the MCADD patient is very favorable [4]. Through early diagnosis of MCADD, the risk of death during derailment reduces to zero and the neurological rest lesions (epilepsy, paralysis, behavioral disorders, developmental disorders) after decompensation are halved [4]. There is thus an important role for preventive medicine where a metabolic disease is transformed into a metabolic disorder by means of simple measures (prevention of fasting and rapid care of sober states) that prevent the development of the disease. These children should be followed in the first 5–7 years of life to avoid decompensation. This can be done by the general practitioner and does not require a specialized center.

1.2. MCADD screening

MCADD in infants can be detected via a blood sample which is taken within a few days after birth using a heel prick test. The heel prick is performed systematically for all newborns in many developed countries (e.g. Denmark, The Netherlands, Germany, Belgium and Luxembourg). The blood sample is subsequently analyzed using tandem mass spectrometry. Depending on the screening center, a derivatized [5] or non-derivatized [6] screening method is used. A sample spectrum is shown in Fig. 1 for both a normal and an MCAD-deficient person.

An increase of the specific acylcarnitine values above established (device-specific) cut-off points usually results in a second blood analysis, carried out when the child is 8 weeks old. This second analysis includes the determination of acylcarnitine values – with tandem mass spectrometry – and the fatty acid profile in plasma. Then the organic acids and glycines in urine are determined [7]. If this second test shows no normalization of the acylcarnitine values, there is need for further review by enzymatic studies and/or DNA analysis.

Many screening centers currently use a derivatized screening method [5]. However, the PCMA as well as some other screening centers in Europe, have switched to using a non-derivatized screening method [6]. The key difference is that the derivatizaton step which requires heating of the dried analytes with dry, acidified (3 N) butanol, is no longer needed. The non-derivatized method requires less processing steps, leads to faster extraction times and has a lower cost for reagents [8].

While both methods show strong correlation among the different measured analytes, it was reported that several analytes showed consistent bias (C0, C2, C10, C16, Gly and Arg) for the non-derivatized method compared to the derivatized method [9]. For four instances this bias was due to higher recovery (C2, C10, C16 and Arg) and for the two others (C0 and Gly) this was due to a lower recovery. This bias can potentially affect the performance of data mining algorithms and may also lead to slightly different models or model parameters compared to models for derivatized data.

1.3. Data mining methods for MCADD classification

Several statistical techniques have been published to establish cutoff values on acylcarnitine values for MCADD classification [10–13]. A comparison of different data mining algorithms for classification of MCADD and other metabolic disorders on derivatized tandem MS neonatal data was done by Baumgartner et al. [14,15]. A feature selection approach for MCADD classification by Ho et al. [16] can be considered as the current state-of-the art data mining method with respect to performance. They reported sensitivity and specificity values of 100% and 99.901% respectively on a dataset of derivatized tandem MS neonatal data in Heidelberg.

This study is the first application of machine learning techniques on *non*-derivatized neonatal screening data. We applied C4.5 decision trees, logistic regression and ridge logistic regression using a grid search approach to optimize model parameter settings, included variables and classification thresholds. Our results using ridge logistic regression show a significantly better performance compared to the current state-of-the-art method for derivatized MS data [16] while our method requires less variable measurements and retains more interpretability.

2. Materials and methods

2.1. Dataset

An anonymized dataset of 44,159 blood samples was collected and analyzed using a non-derivatized tandem MS screening method [6] containing 12 MCADD cases. The dataset consists of two separate parts, each measured using a different screening system.

The *first part* is used as *training data* for learning the classification models. It was obtained as part of a systematic screening for newborns by the PCMA screening center (Belgium) during the first half of 2009. It consists of 32,109 samples and was collected using the *Quattro micro* screening system. This dataset was further enriched with blood samples of all MCADD cases that occurred between 2003 and 2009 at the PCMA screening center, resulting in a total of 9 MCADD samples. These 9 MCADD cases were further confirmed with a genetic test and to our best knowledge, no unidentified MCADD cases are present in the dataset.

The *second part* of the dataset is used as an *independent test set*. It was analyzed using a different screening system (*Xevo QT MS*) and consists of 12,050 samples (collected between June 2010 and September 2010). This dataset contained no MCADD cases from the general population and has been enriched with three spiked blood samples that were provided by the Centre for Disease

Please cite this article in press as: Van den Bulcke T et al. Data mining methods for classification of Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data. J Biomed Inform (2011), doi:10.1016/j.jbi.2010.12.001

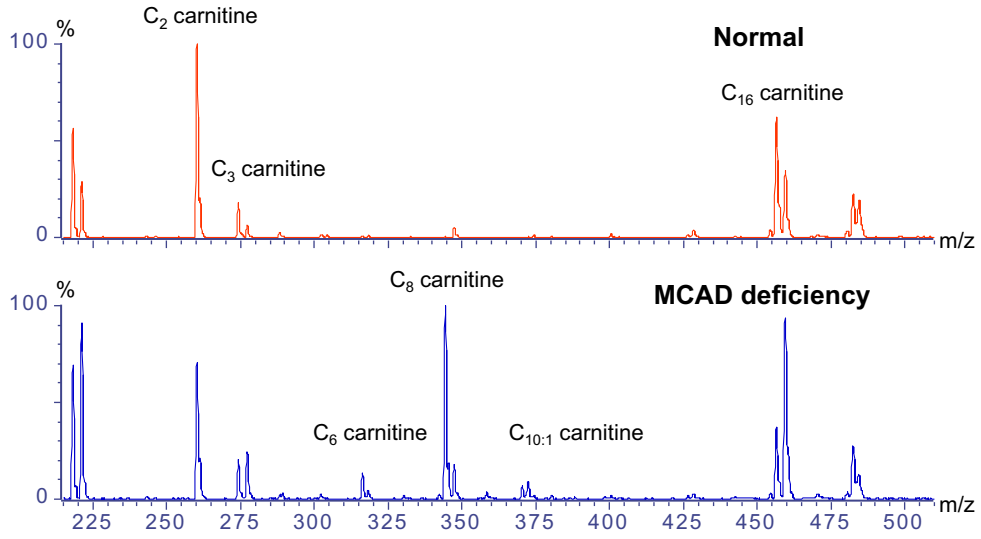

Fig. 1. Tandem MS spectrum for a normal (upper) vs. MCAD deficient (lower) blood sample.

Control (CDC) as external quality controls. These CDC samples are designed to resemble a real MCADD blood sample and any screening center should identify them as positive MCADD cases. While some quantitative differences exist between the CDC samples and real MCADD cases, the most relevant acylcarnitine measurements are comparable. Both sets can therefore be considered sufficiently similar for the presented analyses in this study. We refer to Supplementary File 1 for a detailed comparison between the CDC control samples and real MCADD samples.

The measured parameters for each blood sample are the fatty acid concentrations C0, C3, C5, C5DC, C6, C8, C10, C10:1, C14:1, C16 and concentration ratio's C3/C2, C5DC/C8, C5DC/C16, C8/C2, C8/C10, C8/C12. These measures were further enriched with all possible derived ratios for combinations of the primary concentrations, such as C0/C6, C8/C16, leading to 45 additional variables. All concentrations and ratio's were $\log_{10}$-transformed for the analysis, leading to approximate normal distributions with heavy tails for the concentrations. These measured parameters are the predictor variables and are denoted as X, the binary outcome is denoted as *MCADD* (0 = no MCADD; 1 = MCADD).

2.2. Machine learning methods

Three data mining methods were compared: decision trees, logistic regression and ridge logistic regression. For each of these methods, models were constructed starting from various subsets of the original and derived variables, as shown in Table 1.

In *decision trees*, a (usually binary) classification tree is constructed and a class is assigned to each leaf node. In each internal node, a simple decision rule – usually involving a single variable that is above or below a specific threshold – decides for taking the left or right branch. A decision tree is usually learned from the data by iteratively splitting nodes into child nodes by means of a splitting criterion that maximizes information gain or some other measure that tends to separate the two classes into separate nodes. Decision trees can be overfitted to the data and are therefore usually pruned (i.e. some branches are removed from the tree). One of the most well known and often used algorithms is C4.5 [17], which is used in our experiments.

Binary logistic regression [18] is a widely used statistical technique that constructs a hyperplane between two datasets which separates the two classes. The risk or probability for having MCADD is defined by Eq. (1), where z is defined as a linear function of its predictors x (Eq. (2)). For performing a classification, a cutoff for the probability $f(z)$ is often set at 0.5. In our experimental setup however, this cutoff is also optimized.

$$P(MCADD = true|X = x) = f(z) = \frac{e^z}{e^z + 1} \quad (1)$$

$$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + + \beta_k x_k \quad (2)$$

Ridge logistic regression [19] extends logistic regression by introducing an additional ridge parameter λ in the model. An additional penalty term, $\lambda||\beta||^2$, is added to the log-likelihood where β is the vector of the regression coefficients. For $\lambda = 0$, the model behaves as logistic regression. For larger values of λ, larger model coefficients are penalized and generally tend to be closer to 0. As such, the model coefficients are somewhat biased, whereas the coefficients obtained with normal logistic regression are unbiased estimators of the true coefficients. When the predictor variables are collinear or near collinear, the variances of these coefficients are however much smaller than for logistic regression and the overall error for ridge logistic regression will therefore be lower (see also '*bias-variance tradeoff*' in [20]).

2.3. Experimental setup

The overall experimental setup is shown in pseudo code in Table 2 and is defined as follows. We evaluated three data mining methods (decision trees, logistic regression and ridge logistic regression) for a range of model parameters (such as the ridge parameter for the regression, the confidence threshold for pruning in decision trees etc.) and for different subsets of variables that were included in the model (Table 1).

For each of the combinations model/variable set, an N-fold stratified cross-validation was performed where N is the number of MCADD cases in the dataset. A regular cross-validation would result in an uneven balancing of the MCADD cases over the different folds, therefore a *stratified* N-fold cross-validation was used, resulting in 1 MCADD case for each fold. This procedure was repeated 10 times for different randomizations of the dataset and the results were averaged.

While the main output of the classification algorithms is usually a binary classification (0 = no MCADD; 1 = MCADD), we will use the *probability* of being classified as MCADD as the primary output of the algorithm. This enables us to further improve the performance by choosing an optimal probability threshold value (within the

Please cite this article in press as: Van den Bulcke T et al. Data mining methods for classification of Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data. J Biomed Inform (2011), doi:10.1016/j.jbi.2010.12.001

Table 1
Selected subsets of variables that were evaluated for each of the models and parameter settings. Set 1 includes all measured concentrations and ratios that are used in the newborn screening at the PCMA screening center. Set 2 extends this set by also calculating all other possible ratios between these concentrations and adding them to the variable list. Set 3 includes only measured concentrations, without any ratios. Sets 4–20 include various subsets of variables that are associated with MCADD.

Set	Included variables
1	C0, C3, C3/C2, C5, C5DC, C5DC/C8, C5DC/C16, C6, C8/C2, C8/C10, C8/C12, C10, C10:1, C14:1, C16
2	Set 1 + all possible derived ratio's
3	C0, C3, C5, C6, C5DC, C8, C10, C10:1, C14:1, C16
4	C6, C8, C10, C8/C2
5	C6, C8, C10, C6/C8, C6/C10, C8/C10, C8/C2
6	C0, C6, C8, C10, C8/C2
7	C8, C10, C8/C2
8	C8, C8/C2
9	C8, C8/C2, C6/C8, C6/C10, C8/C10
10	C8, C8/C10
11	C8
12	C8/C2
13	C8, C8/C2, C8/C12
14	C0, C8, C10, C8/C2
15	C0, C8, C8/C2
16	C0, C8, C8/C2, C6/C8, C6/C10, C8/C10
17	C0, C8, C8/C10
18	C0, C8
19	C0, C8/C2
20	C0, C8, C8/C2, C8/C12

cross-validation loop) to maximize the sensitivity such that all MCADD cases are correctly identified while keeping the number of false positives minimal.

A straightforward choice of this threshold would be to select the threshold value where all MCADD cases in the *training set* are identified as positive and the number of false positives in the training set is minimal, called $threshold_{TRAIN}$. This threshold value is however likely to be too conservative and will fail to identify some MCADD cases in the *test* set. In order to accommodate for this, an adjustment factor is defined that introduces an additional margin on this threshold, which is illustrated in Fig. 2. All predicted probabilities $P(\text{MCADD} = true|X = x)$ are sorted and $threshold_{TRAIN}$ is the probability threshold value such that all actual MCADD cases are equal or above this threshold. We define K as the *total* number of cases that is equal or above this threshold. This index K is multiplied with an adjustment factor F_{ADJ}, leading to a new threshold $threshold_{OPT}$. This is the $(F_{ADJ} \cdot K)$th element in the sorted $P(\text{MCADD} = true|X = x)$ vector, where $(F_{ADJ} \cdot K)$ is rounded to the nearest integer.

The following performance measures were used to evaluate the models: $sensitivity = TP/(TP + FN)$, $specificity = TN/(TN + FP)$ and *positive predictive value* $(PPV) = TP/(TP + FP)$ where TP indicates the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. As explained below, calculations of PPV were corrected for the true MCADD prevalence.

3. Results

The models were evaluated for a broad range of adjustment factors (0.5–5) and model parameter settings. The best models together with their associated parameter settings are shown in Table 3. Model A is the best decision tree model; model B shows the best model for logistic regression with only C8 as a predictor variable; model C is the best model with logistic regression for all possible variable sets; and model D is the overall best model, namely a ridge logistic regression model.

Previously reported performance measures for MCADD prediction (Baumgartner et al. [14]); Ho et al. [16]) do not accurately reflect the expected performance on datasets of *real populations*: both studies calculate performances on datasets which are enriched with MCADD cases (or which are depleted with non-MCADD cases), leading to highly biased performance measures such as the positive predictive value (PPV). The PPV was therefore recalculated to assess how many false positives we would encounter in a *real* population. In order to allow a fair comparison, an MCADD prevalence of 1/15,000 is used for all studies to recalculate the PPV. Because the training data was highly enriched for MCADD cases, the calculated PPV values for a *real* population are significantly different from the reported PPV values in the respective publications. Since the dataset enrichment has no effect on sensitivity and specificity measures, no recalculation of these measures was required. The performance measures shown in Table 3 are derived from the cross-validation analysis on the *training data* in the respective studies. The reason is that both studies included part of the training data in their test sets and the reported results may therefore be slightly too optimistic for the test data. Using the training data and recalibrating the results more accurately reflects the true expected performance.

3.1. Decision tree models

Firstly, a number of decision tree models were evaluated. The best performing decision tree model was *model A*, which achieves high sensitivity (98.889%) but does not identify all MCADD cases in a cross-validation setting. The selected variables in the model were C8, C10 and C8/C2, which are known to be indicative for MCADD. The decision tree model of Baumgartner et al. [14], *model E*, also included C8 and C10 but excluded the C8/C2 ratio in favor of the C16 concentration. The PPV of our best model (19.21%) is significantly higher than the current state of the art decision tree model [14] (1.30%). The differences in performance (PPV) between our method and current state-of-the-art could be either due to the nature of the dataset, due to the different methodology or a combination of both (see Section 4). From a medical perspective, still an unacceptably high number of MCADD patients are diagnosed as healthy using the decision tree method (>1%). Decision trees were also less robust w.r.t. variations in parameter settings and variable selection choices, leading to larger variations in sensitivity and specificity compared to logistic regression for the range of evaluated models (results not shown). These two undesired properties make decision trees less appropriate choices for modeling MCADD.

3.2. Logistic regression analysis using only C8

Secondly, a series of logistic regression analyses were performed with one single parameter C8, the predominant metabolite associated with MCADD. The best model, namely *model B* (Table 3), resulted in good predictions: 100% of the MCADD cases were identified in the cross-validation setting with an average specificity of 99.965% over the different randomizations. Comparing the results with a normal logistic regression (without threshold optimization) for the same dataset (*model H*), we see that our threshold optimization strategy increases sensitivity from 62.5% to 100%, while decreasing the PPV (22.9–15.9%) and specificity (99.988–99.965%). This result confirms that statistical models that only use C8 as a predictor can already achieve a high classification accuracy [14]. This is also illustrated in Fig. 3 which shows the log-concentration of C8 for MCADD (red diamonds) and normal (black dots) cases.

Please cite this article in press as: Van den Bulcke T et al. Data mining methods for classification of Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data. J Biomed Inform (2011), doi:10.1016/j.jbi.2010.12.001

Table 2
Pseudo code of the experimental setup.

```
for each data mining method M:
  for each set of model parameters P:
    for each set of variables V:
      for each randomization R:
        for each cross-validation-fold F:
          testData = rowSubset(dataset, select = fold F, rand = R)
          trainDataAllColumns = rowSubset(dataset, select = all except fold F, rand = R)
          trainData = colSubset(trainDataAllColumns, V)
          model = train(M, trainData)
          trainProbabilities[F] = predict(model, trainData)
          thresholdTrain = calcThresholdTrain(trainProbabilities[F], trainData)
          testProbabilities[F] = predict(model, testData)
          for each adjustment factor F_ADJ:
            measures[M, P, V, R, F, F_ADJ] = calcMeasures(testProbabilities, F_ADJ * threshold_TRAIN)
measuresAggr = aggregate(measures, F, type = sum) # sum the results overall CV-folds
measuresAvg = aggregate(measuresAggr, R, type = mean) # average over the different randomizations
modelA = select (M, P, V, FADJ) for which
  measuresAvg[k].M() = = "C4.5" and
  measuresAvg.specificity() is maximal and
  measuresAvg.sensitivity() = = max(measuresAvg[M = "C4.5"].sensitivity())
modelB = select (M, P, V, FADJ) from measuresAvg[k] for which
  measuresAvg[k].specificity() is maximal and
  measuresAvg[k].sensitivity() = = max(measuresAvg[M = "Logistic", P = "λ = 0", V = "C8"].sensitivity())
modelC = select (M, P, V, FADJ) from measuresAvg[k] for which
  measuresAvg[k].M() = = "Logistic" and measuresAvg[k].P() = = "λ = 0" and
  measuresAvg[k].specificity() is maximal and
  measuresAvg[k].sensitivity() = = max(measuresAvg[M = "Logistic", P = "λ = 0"].sensitivity())
modelD = bestModel = select (M, P, V, F_ADJ) from measuresAvg[k] for which
  measuresAvg[k].specificity() is maximal and
  measuresAvg[k].sensitivity() = = max(measuresAvg.sensitivity())
```

3.3. Logistic regression analysis using all variable sets

The positive predictive value and specificity can be further increased by including more variables in the regression model without reducing the sensitivity. The best performing logistic regression model for *all* of the variable sets (without using ridge regression), is *model C* (Table 3). A number of additional variables to C8 are included, namely C8/C2, C6/C8, C6/C10 and C8/C10, that are known to be predictive for MCADD [16]. By including these variables, the (recalibrated) PPV is increased from 15.92% to 23.41% while retaining 100% sensitivity.

3.4. Ridge logistic regression

The overall best model, namely *model* D, uses ridge logistic regression [19] and includes the same set of variables as *model C*. Compared to logistic regression, the recalibrated PPV was further increased to 33.90% while the sensitivity remained at 100%, thereby outperforming the current state-of-the-art methods on derivatized datasets. The PPV for our independent test dataset (28.67%) confirms that this high performance is not due to overfitting. By using ridge regression, the contribution of less predictive variables and highly correlated variables in the model are reduced, thereby reducing overfitting of the model.

4. Discussion

In this study, the performance of a number of data mining methods were compared for predicting MCADD classification based on acylcarnitine measurements. In situations where there is a large asymmetry between misclassifying a positive vs. a negative sample, cost-based classifiers [21] are frequently used which optimize a loss function rather than minimizing the overall misclassification rate. While this approach is certainly applicable, it was not used in this study. Rather we have chosen to directly optimize the probability cutoff, which is more directly linked to our goal of maximizing the sensitivity.

The current state of the art method for the *derivatized* screening method was developed by Ho et al. [16] and is shown as model F in Table 3. The model by Ho et al. that is based on derivatized screening data (model E, Table 3) uses seven different variables, namely C4, C5:1, C8, C10, C14OH, C16OH and C18:1. While the model achieves 100% sensitivity, the expected positive predictive value for a real population is low (6.3%). The model uses complex features to classify MCADD, as shown in Eq. (3), and is therefore slightly more difficult to interpret. The best performing model in our analysis requires less variable measurements, namely 5, and achieves a strongly improved performance (sensitivity = 100%; PPV = 33.90%).

$$(C8 - C4 + C5:1 + C10) > 0.004 \text{ and } \frac{C8 + C18:1}{C14OH + C16OH} > -2.012 \quad (3)$$

As stated, there are quantitative differences between the measurements for the derivatized screening and non-derivatized screening methods which may affect the performance of the data mining methods. The differences in performance (PPV) between our method and current state-of-the-art could be due to the nature of the dataset (non-derivatized), due to the different methodology or due to a combination of both. It is difficult to answer this question conclusively. We do observe that our threshold optimization method strongly improves performance of logistic regression on the same non-derivatized dataset (Table 3, *model H* vs. *model B*). This indicates that at least part of the performance difference between our models and current state-of-the-art on derivatized datasets can be explained due to our methodology.

Some remarks are necessary regarding the optimal value for the adjustment factor F_{ADJ}. Because we have only a small sample (nine cases) of the full distribution of MCADD cases, we do not have sufficient measurements from the tails of this full distribution. Our

Please cite this article in press as: Van den Bulcke T et al. Data mining methods for classification of Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data. J Biomed Inform (2011), doi:10.1016/j.jbi.2010.12.001

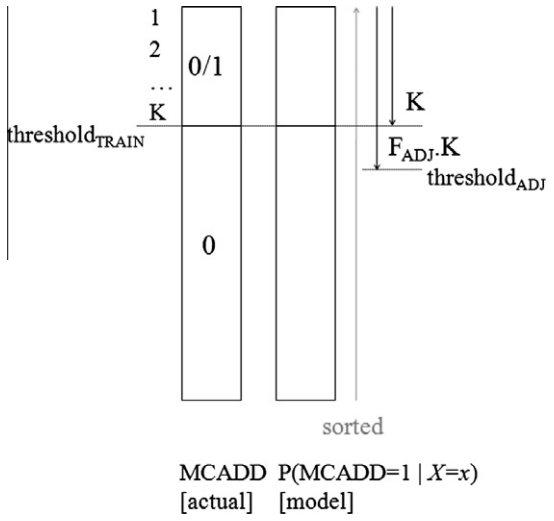

Fig. 2. Illustration of the calculation of the adjusted threshold $threshold_{ADJ}$.

confidence threshold is estimated based on maximizing the expected sensitivity of the dataset but it does not account for the variance on this estimate. This implies that, while the *estimate* on the sensitivity is 100%, the real sensitivity may deviate from this number. As such, an analysis on a larger dataset may lead to a lower specificity or sensitivity. Unfortunately, due to the small number of MCADD samples, a standard deviation on this estimate cannot be reliably estimated. In order to use these models in practice, either a larger dataset is required with more MCADD cases to train the model or an additional safety margin for the adjustment factor can be chosen.

Our approach of threshold optimization in combination with ridge logistic regression leads to very stable performances, independent of the exact value of ridge regression coefficient λ for a broad range of values. The reason for this behavior is likely because of the threshold optimization method that compensates any change in predicted probabilities by adjusting the optimal threshold value. In general, ridge logistic regression combined with threshold optimization has desirable properties for practical applications. Firstly, it leads to an interpretable model since a (logistic) regression model is used. Secondly, the performance is stable w.r.t. large variations in parameter settings. Thirdly, the adjustment factor can easily be interpreted and adjusted by a domain expert, e.g. to be more conservative than strictly required.

Currently, one of the main difficulties in modeling MCADD and other metabolic disorders, is that these models cannot be directly used by other screening centers due to variation between screening centers in equipment and sample preparation. Moreover, each new batch of sampling kits in a screening center requires a recalibration of the experimental setup. It would be desirable, both from a medical and data mining perspective, to establish a gold standard across screening centers that enables a universal calibration of the screening results, both for non-derivatized and derivatized methods. This would ensure the direct interchangeability of statistical models and would also allow researchers to combine datasets from different screening centers in order to leverage the information contained in each of the individual ones. This is especially important for rare diseases as the number of cases in each screening center is relatively low.

If mathematical models would be used in practice to support diagnostic decision making, several requirements need to be met. In case of MCADD, no false negatives should occur as experts are currently able to identify all MCADD cases. A decision support system that would fail to identify an MCADD case is therefore unacceptable from a medical point of view. In that respect, decision trees seem to have less interesting properties for predicting MCADD. Their higher susceptibility to parameter settings and random variations in the dataset, combined with a sensitivity well below 100% make them a less reliable method compared to logistic regression in the context of MCADD classification. Ensemble methods such as random forests could possibly reduce these undesirable properties, however at the cost of interpretability of the model.

Table 3
Performance measures for the best performing models. The best performing models are shown in models A–D: the best decision tree model (A); the best logistic regression model (no ridge) with only C8 (B); the best logistic regression model (no ridge) (C); and the overall best model, namely a ridge logistic regression model (D). The variable set shows the variables that were selected by the model. For models A–D the original variable set that was used for training the model, is also shown between brackets (see also Table 1). The sensitivity, specificity and positive predictive value indicate the performance in a cross-validation setting for all models, which is the expected performance on unseen data. The column *cohort size* reflects the total size of the dataset, while *training data* shows the actual size of the dataset used for training the models (values in brackets indicate the number of MCADD cases for both columns). The bold PPV values are derived from cross-validation analyses, the values in italics for models A–D are measured on our independent test set. For reference, the models of Baumgartner et al. [14] and Ho et al. [16] are shown as current state-of–the-art for decision trees (*E*), logistic regression (*F*) and feature construction (*G*) methods, based on derivatized tandem MS data. Model H shows the results for a normal logistic regression *without* threshold optimization on the non-derivatized dataset of the PCMA. Note that the reported numbers of models E–G are not directly comparable to those of models A–D, as different datasets were used for the analyses. Used abbreviations: *PT*: confidence threshold for pruning [17], *M*: minimal number of instances per leaf, F_{ADJ}: adjustment factor, λ: ridge parameter of the logistic regression.

Model		Selected variables	Screening technique	Cohort size	Training data	Sensit.(%)	Specif.(%)	PPV*(%)
[A]	C4.5 decision trees – OPT (PT = 0.25, M = 2, F_{ADJ} = 1.01)	C8, C10, C8/C2 (from set 5)	non-deriv.	44159 (12)	32109 (9)	98.889	99.972	19.21 (15.16)
[B]	Logistic regression (C8 only) – OPT (λ = 0, F_{ADJ} = 1.05)	C8 (from set 11)	non-deriv.	44159 (12)	32109 (9)	100.000	99.965	15.92 (16.73)
[C]	Logistic regression – OPT (λ = 0, F_{ADJ} = 1.50)	C6/C8, C6/C10, C8, C8/C2, C8/C10 (from set 5)	non-deriv.	44159 (12)	32109 (9)	100.000	99.978	23.41 (16.73)
[D]	Ridge logistic regression – OPT (λ = 0.001, F_{ADJ} = 1.15)	C6/C8, C6/C10, C8, C8/C2, C8/C10 (from set 5)	non-deriv.	44159 (12)	32109 (9)	100.000	99.987	33.90 (28.67)
[E]	C4.5 decision trees Baumgartner et al. [16]	C8, C10:1, C16	derivatized	590466 (63)	1241 (63)	95.238	99.517	1.30
[F]	Logistic regression Baumgartner et al. [16]	C8	derivatized	590466 (63)	1241 (63)	95.238	99.839	3.79
[G]	Feature constr. (quantile variant) Ho et al. [18]	C4, C5:1, C8, C10, C14OH, C16OH, C18:1	derivatized	397195 (30)	1685 (30)	100.000	99.901	6.31
[H]	Logistic regression (C8 only)	C8	non-deriv.	44159 (12)	32109 (9)	62.500	99.988	22.91

*The reported positive predictive value (PPV) is adjusted for the enrichment in the datasets, such that it represents the PPV of an actual population with an MCADD prevalence of 1/15,000. Note that sensitivity and specificity are identical for enriched and normal datasets and thus require no adjustment.

Please cite this article in press as: Van den Bulcke T et al. Data mining methods for classification of Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data. J Biomed Inform (2011), doi:10.1016/j.jbi.2010.12.001

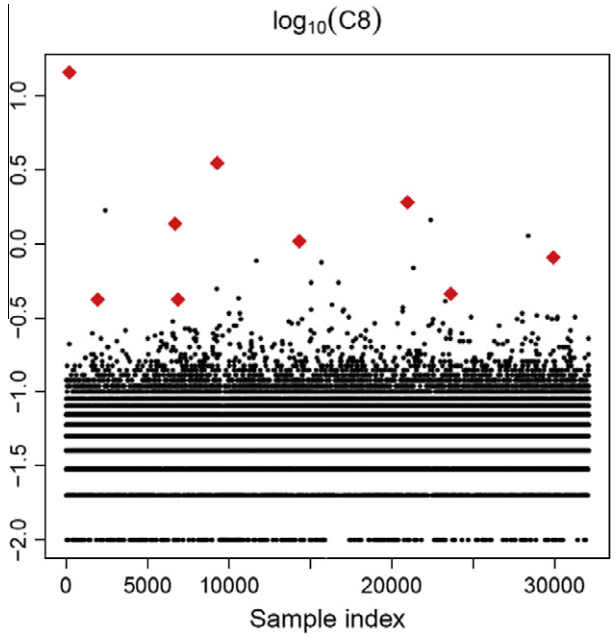

Fig. 3. Log-concentration of C8 acylcarnitines for MCADD cases (red diamonds) and normal cases (black circles). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Conclusions

We have applied the first data mining methods for modeling MCAD deficiency using *non-derivatized* tandem MS datasets. Three different data mining methods were evaluated for a broad range of model parameters, namely decision trees, logistic regression and ridge logistic regression. The best performing model is a ridge logistic regression model that shows a significantly better performance compared to the current state of the art for the *derivatized* screening method while retaining more interpretability and requiring a lower number of acylcarnitine measurements. A sensitivity of 100%, a specificity of 99.987% and a positive predictive value of 33.90% (recalibrated for a real population) were achieved in a stratified cross-validation setting for non-derivatized screening data, outperforming current state-of-the-art methods for MCADD prediction on derivatized data. Our analysis was performed on a relatively small dataset of 44,159 cases with 12 MCADD cases. A further confirmation of these results on larger derivatized and non-derivatized screening datasets is therefore desired. The results indicate the potential value of data mining methods as a diagnostic support tool and show that strong classification performances are achieved for non-derivatized tandem MS data using a method that combines ridge logistic regression with variable selection and threshold optimization.

Conflict of interest

Nothing to report.

Authors contributions

Conception and design: FE, TVDB, PVB; provision of study material: FE, PVB; data analysis and interpretation: TVDB, PVB, KW; manuscript writing: TVDB, PVB, FE, VVH, TVG, SP; final approval: TVDB, PVB, VVH, SVB, KW, GS, ES, SP, TVG, FE.

Acknowledgments

This research was funded by University Hospital Antwerp and Provinciaal Centrum voor de Opsporing van Metabole Aandoeningen (PCMA). The PCMA has kindly provided us with the screening datasets. We would like to thank the reviewers for their most valuable feedback and suggestions.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2010.12.001.

References

[1] Fischer A, Borensztein P, Roussel C. The European rare diseases therapeutic initiative. PLoS Med 2005;2(9).
[2] Kompare M, Rizzo W. Mitochondrial fatty-acid oxidation disorders. Semin Pediatr Neurol 2008;15(3):140–9.
[3] Wilcken B. Fatty acid oxidation disorders: outcome and long-term prognosis. J Inherit Metab Dis 2010.
[4] Van Hove JL, Zhang W, Kahler SG, Roe CR, Chen YT, Terada N, et al. Medium-chain acyl-CoA dehydrogenase (MCAD) deficiency: diagnosis by acylcarnitine analysis in blood. Am J Hum Genet 1993;52(5):958–66.
[5] Chace D, Hillman S, Van Hove J, Naylor E. Rapid diagnosis of MCAD deficiency: quantitative analysis of octanoylcarnitine and other acylcarnitines in newborn blood spots by tandem mass spectrometry. Clin Chem 1997;43(11):2106–13.
[6] Nagy K, Tákats Z, Pollreisz F, Szabó T, Vékey K. Direct tandem mass spectrometric analysis of amino acids in dried blood spots without chemical derivatization for neonatal screening. Rapid Commun Mass Spectrom 2003;17(9):983–90.
[7] Matern D, Rinaldo P. Medium-chain acyl-coenzyme a dehydrogenase deficiency. PLoS Genet 2005.
[8] Eyskens FJM, Philips E. Newborn mass screening using tandem mass spectrometry: results of the validation and comparison of two methods (derivatized/non-derivatized). J Inherit Metab Dis 2007;30(Suppl. 1) [3].
[9] http://las.perkinelmer.com/content/RelatedMaterials/Posters/SPS_MSMSComparison.pdf [accessed 05.07.10].
[10] Chace D, Kalas T, Naylor E. Use of tandem mass spectrometry for multianalyte screening of dried blood specimens from newborns. Clin Chem 2003;49(11):1797–817.
[11] Clayton P, Doig M, Ghafari S, Meaney C, Taylor C, Leonard J, et al. Screening for medium chain acyl-CoA dehydrogenase deficiency using electrospray ionisation tandem mass spectrometry. Arch Dis Child 1998;79(2):109–15.
[12] Pourfarzam M, Morris A, Appleton M, Craft A, Bartlett K. Neonatal screening for medium-chain acyl-CoA dehydrogenase deficiency. Lancet 2001;358(9287):1063–4.
[13] Okun J. A method for quantitative acylcarnitine profiling in human skin fibroblasts using unlabelled palmitic acid: diagnosis of fatty acid oxidation disorders and differentiation between biochemical phenotypes of MCAD deficiency. Biochim Biophys Acta 2002;1584(2–3):91–8.
[14] Baumgartner C, Böhm C, Baumgartner D. Modelling of classification rules on metabolic patterns including machine learning and expert knowledge. J Biomed Inform 2005;38(2):89–98.
[15] Baumgartner C, Bohm C, Baumgartner D, Marini G, Weinberger K, Olgemoller B, et al. Supervised machine learning techniques for the classification of metabolic disorders in newborns. Bioinformatics 2004;20(17):2985–96.
[16] Ho S, Lukacs Z, Hoffmann G, Lindner M, Wetter T. Feature construction can improve diagnostic criteria for high-dimensional metabolic data in newborn screening for medium-chain acyl-CoA dehydrogenase deficiency. Clin Chem 2007;53(7):1330–7.
[17] Quinlan R. C45: programs for machine learning. Morgan Kaufmann; 1993.
[18] Hosmer D, Lemeshow S. Applied logistic regression. Wiley-Interscience Publication; 2000.
[19] Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. Appl Stat 1992;41(1):191–201.
[20] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Data mining, inference, and prediction. Springer; 2009.
[21] Domingos P. MetaCost: a general method for making classifiers cost-sensitive. KDD. 155-164. San Diego, California, United States: ACM; 1999.